The influence of candidates' physical attributes on assessors' ratings in clinical practice

Article (Published Version)

Medical Teacher

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/imte20

The influence of candidates' physical attributes on assessors' ratings in clinical practice

A. H. Sam, M. D. Reid, V. Thakerar, M. Gurnell, R. Westacott, P. Yeates, M. W. R. Reed & C. A. Brown

To cite this article: A. H. Sam, M. D. Reid, V. Thakerar, M. Gurnell, R. Westacott, P. Yeates, M. W. R. Reed & C. A. Brown (2021) The influence of candidates' physical attributes on assessors' ratings in clinical practice, Medical Teacher, 43:5, 554-559, DOI: 10.1080/0142159X.2021.1877268

To link to this article: https://doi.org/10.1080/0142159X.2021.1877268

Published online: 11 Feb 2021.

MEDICAL TEACHER
2021, VOL. 43, NO. 5, 554–559
https://doi.org/10.1080/0142159X.2021.1877268

The influence of candidates' physical attributes on assessors' ratings in clinical practice

A. H. Sam[a], M. D. Reid[a], V. Thakerar[a], M. Gurnell[b], R. Westacott[c], P. Yeates[d,e], M. W. R. Reed[f] and C. A. Brown[g]

[a]Imperial College School of Medicine, Imperial College London, London, UK; [b]Wellcome–MRC Institute of Metabolic Science, University of Cambridge and NIHR Cambridge Biomedical Research Centre, Cambridge University Hospitals, Cambridge, UK; [c]Birmingham Medical School, University of Birmingham, Birmingham, UK; [d]School of Medicine, Keele University, Keele, UK; [e]Fairfield General Hospital, Pennine Acute Hospitals NHS Trust, Bury, UK; [f]Brighton and Sussex Medical School, University of Sussex, Brighton, UK; [g]Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, UK

ABSTRACT

Background: Assessments of physician competence in the work-place are common and often contribute to high-stakes assessments. Previous research suggests that assessors' judgements can be influenced by candidates' physical attributes. We investigated whether simulated candidates' scores were influenced by assessor bias based on tattoos, hair colour, and a regional accent.

Methods: We used an experimental, video-based, single-blinded, randomised, internet-based design. We created videos of simulated medical intern performances of a clinical examination at four different standards of competence. Four videos were also created of simulated candidates performing at a 'clear pass' standard, with either no stereotypical attribute (CPX), purple hair (CPH), tattoos (CPT) or a Liverpool English accent (CPA). Assessors were randomly assigned to watch five videos including the "clear pass" candidate without an attribute and one of the "clear pass" candidates with an attribute and asked to give an overall global grade for each candidate. We compared the global grades for the clear pass candidates with and without attributes.

Results: Ninety-eight assessors were included in the analysis. The total scores for the candidates with stereotyped attributes were not significantly lower than the candidate with no attribute. Assessors showed moderate levels of agreement between the global grades awarded for all the candidates. The global grades awarded to candidate with a stereotypical attribute were not significantly lower than for those without.

Conclusions: The presence of tattoos, purple hair, or a regional accent did not systematically negatively influence the grade or score awarded by assessors to candidates in observed clinical examination scenarios.

KEYWORDS
Assessment; bias; medicine; clinical

Introduction

Ratings based on observations of a physician's competence in practice by senior colleagues occur frequently and have traditionally contributed to learning in the workplace as part of an apprenticeship model (Swanwick 2005). Workplace-based assessments of competence, such as the mini-clinical evaluation exercise (mini-CEX), have been generally supported (Hatala et al. 2006; Norcini and Burch 2007) and are increasingly being integrated into postgraduate curricula across the world (Miller & Archer 2010). However, concerns have been raised about the validity and reliability of such methods and their use as part of high-stakes assessments (Hawkins et al. 2010). It is well established that assessors are prone to variability due to cognitive biases such as leniency, inconsistency, and the halo effect (McManus et al. 2006; Iramaneerat and Yudkowsky 2007; Harasym et al. 2008). Individual examiners have also been shown to rely on value-based judgements which are prone to stereotype bias (Williams et al. 2003). Attempts to reduce the impact of these sources of assessor variability have shown limited effect (Cook et al. 2009) such that they may ultimately threaten the validity and objectivity of the assessment format (Hawkins et al. 2010). This paper contributes to the developing understanding of sources of assessor variability due to bias.

Practice points

- Assessments of competence based on observations of practice are common in healthcare settings.
- Individual assessor bias based on candidate characteristics has been previously documented.
- Systematic bias based on hair colour, tattoos, and UK regional accent does not seem to negatively impact the scores or grades awarded by assessors when rating competent candidates.

CONTACT C. A. Brown Celia.Brown@warwick.ac.uk Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, CV4 7HL, UK
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

The role of assessor inferences about candidate attributes such as body language, accent and appearance has been shown to contribute to ratings (Kogan et al. 2011), but have not been explored in a large-scale study. Further work regarding the origins of assessor variability in direct observation assessments has resulted in a proposed model of 'information integration' by assessors which describes the formation of a general impression of a candidate first, followed by the generation of domain scores second, rather than the reverse process which is the intended method of such systems (Yeates et al. 2013). Previous studies have shown that some physical attributes such as an individual's ethnicity have an impact on their attainment in both undergraduate and postgraduate medical examinations (Woolf et al. 2011). This effect may be partly attributed to bias on behalf of the assessors but its overall origins are not clear (Yeates et al. 2013). It is also apparent that amongst physicians in clinical practice bias based on ethnicity persists and contributes to healthcare disparity for patients (Stone and Moskowitz 2011; Dovidio and Fiske 2012; Moskowitz et al. 2012).

Stereotypes amongst the general population about those with tattoos (Wohlrab et al. 2007), extremes of hair colour (Beddow 2011) and accents (Gluszek and Dovidio 2010) are widespread. In particular, Liverpool English accents have been shown to be perceived as less trustworthy than Standard Southern British English (SSBE) (Torre et al. 2018) and lower in prestige and social attractiveness (Bishop et al. 2005). Activation of these stereotypes has been shown to have an impact on real-world outcomes such as success in job interviews, average salary and perceived professionalism (Johnston 2010; Deprez-Sims and Morris 2010; Ruetzler et al. 2012), but there has been no work done to explore their role in assessment within healthcare professionals' education. Despite best efforts, physicians remain prone to the same implicit biases as the general population which may unconsciously impact decision making (Chapman et al. 2013). In some cases bias based on these stereotypes is more overt, such that physicians have been shown to openly express a preference for their colleagues to be dressed according to established norms and where individuals deviate from this standard peers may perceive this as a professionalism concern (Gjerdingen et al. 1987). Stereotypes are more likely to be activated and result in bias when judgements are mentally demanding (Macrae et al. 1994), for example during medical exams (Tavares and Eva 2014). Physical attributes therefore present a potential source of bias that may influence assessor ratings and challenge the validity of workplace-based assessments of competence in clinical practice. This study therefore sought to establish whether the presence of a variety of physical attributes amongst candidates performing at a standardised level had any effect on assessors' ratings.

Methods

Study design

We used an experimental, video-based, single-blinded, randomised, internet-based design.

Procedure

Seven 10 min videos were created of simulated candidates completing a clinical examination typical of those observed in practice (a cranial nerve examination). Volunteer Clinical Teaching Fellows affiliated with Imperial College London were recruited for this role. All simulated candidates were female, of white ethnicity, and a similar age to avoid potential confounding based on these factors. Four of the videos demonstrated the simulated candidates performing the examination at one of four overall performance levels: 'clear fail' (CF), 'borderline' (BD), 'clear pass' (CPX) or 'good' (GD). The other three videos showed a candidate performing at a 'clear pass' level but with either purple hair (CPH), tattoos (CPT), or a Liverpool English accent (CPA). The simulated candidates in all videos except CPA performed with a SSBE accent. Each candidate followed a script created by a panel of experienced examiners to ensure they were performing at the appropriate level and to standardise those performing at the 'clear pass' level. Twelve sets of five videos were then created; with every set including a video of a candidate performing at each of the overall performance levels as well as one video of a candidate with a physical attribute performing at a 'clear pass' level (Appendix 1). The ordering of the five videos differed across the 12 sets to mitigate any bias associated with ordering effects. Each participant was randomly allocated to one of the 12 video sets.

Recruitment and consent

The study was approved by the Medical Education Ethics Committee at Imperial College London (MEEC1718-105). Each medical school in the UK was contacted via the Medical Schools Council and invited to take part in the study. Heads of assessment at each medical school were encouraged to invite a representative sample of assessors to participate in the study via the study website. Participants were informed that they were taking part in a study exploring inter-rater reliability amongst assessors but were not informed that the study aimed to evaluate the impact of physical attributes on scores and performance levels. No identifiable information was collected about the participants. Participants were required to be clinicians with at least one prior experience of formally assessing medical students in clinical examinations. Participants were informed that completion of the marksheets for all five videos and submission of the post-completion questionnaire was evidence of consent. Participants were able to withdraw from the process by closing the web browser at any time prior to completion of the study but due to the lack of collection of identifiable data, were not able to withdraw after submitting their results. Any incomplete data, where participants did not view and score all five videos, were not used in the analysis.

Measures

Participants were asked to assess the candidates at the level expected of a foundation year 1 doctor (equivalent to a medical intern). Participants viewed the five videos and were provided with a blank mark sheet to complete alongside each video (Figure 1). Participants marked each candidate in four domains; 'Physical examination', 'Identify physical signs and

Mark Sheet: Cranial Nerve Examination

Domain 1. Physical examination
Task: Examines the cranial nerves (I-XII)
Excellent (4)
Good (3)
Adequate (2)
Fail (1)
Severe fail (0)

Domain 2. Identifying physical signs and the most likely diagnosis
Task: Reports abnormal findings and offers the most likely diagnosis
Excellent (4)
Good (3)
Adequate (2)
Fail (1)
Severe fail (0)

Domain 3. Clinical management skills
Task: Explains management of patient
Excellent (4)
Good (3)
Adequate (2)
Fail (1)
Severe fail (0)

Domain 4. Interpersonal skills
Task: Communicates appropriately with the patient and examiner
Excellent (4)
Good (3)
Adequate (2)
Fail (1)
Severe fail (0)

Global Grade
Good
Clear Pass
Borderline
Fail

Figure 1. Sample mark sheet.

the most likely diagnosis', 'Clinical management skills', and 'Interpersonal skills'. Each domain was scored between 0−4, with a maximum possible total score of 16. Participants were also asked to assign each candidate a global grade of either 'clear fail', 'borderline', 'clear pass' or 'good'. Participants were able to return to mark sheets for previous candidates but were not able to pause, rewind or replay the videos, to reflect the contemporaneous nature of rating a competency in practice. Following completion of the mark sheets for all five videos, participants were asked to confirm their assessment experience, job role, gender, ethnicity and the geographical region where they worked.

Statistical analysis

Data management and analysis were conducted using Stata V16. Total scores and global grades for each candidate with an attribute were compared with those of the clear pass candidate without an attribute. Individual Wilcoxon matched-pairs signed-ranks tests were used to compare the total scores. Weighted kappa analysis was used to compare the global grades, followed by a Wilcoxon analysis to measure the direction of any disagreement. A p-value of less than 0.0167 was required for statistical significance to account for multiple comparisons within each type of score.

Results

Participants

One-hundred and twenty assessors participated in the study, of whom 98 were included in the analysis (five assessors were removed due to a self-reported lack of experience; seventeen participants did not complete viewing and rating all of the candidates). Table 1 shows the demographic details of all participants included in the analysis. Participants included in the analysis came from ten distinct regions across the UK. Participants were varied in their level of experience and job role. The number of participants who viewed and rated each of the 12 sets of videos was comparable.

Total score

Total scores for all the clear pass candidates ranged from 8 to 16/16 (median 14, interquartile range [IQR] 12 to 15. Figure 2). The modal scores for each candidate were as follows; CPX = 14, CPH = 16, CPT = 12, CPA = 14. Individual Wilcoxon matched-pairs signed-ranks tests were performed on the total scores for the clear pass candidate with no attribute when compared to each clear pass candidate with an attribute. For the candidate with purple hair (CPH) this indicated that scores were statistically significantly higher (median paired difference 1, range −4 to 8, $Z = 2.42$, $p = 0.01$). There was no significant difference between the total scores for CPX and CPT (median paired difference 1, range −4 to 3, $Z = 1.68$, $p = 0.09$) or between CPX and CPA (median paired difference −1, range −3 to 2, $Z = 1.26$, $p = 0.22$)

Global grade

Global Grades for all candidates varied from borderline to good. A weighted kappa analysis using linear weights showed individual assessors had moderate agreement between the global grades awarded to CPX and CPT ($K = 0.412$, $p = 0.007$) and to CPA ($K = 0.446$, $p = 0.004$). There was no significant agreement between the global grade awarded to CPH when compared to CPX ($K = 0.158$, $p = 0.129$). A Wilcoxon matched-pairs signed-ranks analysis was performed to measure the direction of this difference by applying numerical values to each global grade, where fail = 1, borderline = 2, clear pass = 3 and good = 4. This showed no statistically significant difference ($Z = 2.13$, $p = 0.06$) but confirmed the median paired difference for CPH was 1 grade higher than for CPX. Figure 3 shows the number of assessors giving each configuration of global grades to each candidate.

Discussion

For the first time we have compared the influence of hair colour, tattoos and accent on the ratings clinicians give to simulated performances of clinical examinations by candidates. There was no negative impact on the global grades awarded by assessors despite the presence of stereotyped physical attributes. Similarly, the total scores for clear pass candidates with physical attributes were not significantly lower than for the candidate without these characteristics. Interestingly, assessors gave higher total scores and global grades to the candidate with purple hair than to the candidate performing at the same level without a physical attribute. These findings are largely reassuring and suggest that any assessor bias based on the presence of tattoos, hair colour and accent does not negatively influence their judgement. This finding

Table 1. Participant descriptives for all participants, and for the participants who rated the performance of the candidates performing at a 'clear pass' level who also had the presence of a physical attribute.

All participants $N = 98$			CPH $n = 32$		CPT $n = 34$		CPA $n = 32$	
Demographics	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)
Experience								
None	0	0	0	0	0	0	0	0
1–2 Exams	6	6.12	2	6.25	1	2.94	3	9.38
3–4 Exams	13	13.27	5	15.63	6	17.65	2	6.25
5+ Exams	79	80.61	25	78.13	27	79.41	27	84.38
Job role								
Consultant	40	40.82	9	28.13	17	50	14	43.75
Primary Care Physician	33	33.67	0	0	0	0	0	0
Specialty Training years 3+	1	1.02	1	3.13	0	0	0	0
Core Training or Specialty Training years 1–2	1	1.02	0	0	0	0	1	3.13
Other/please specify role & grade if appropriate	22	22.45	6	18.75	9	26.47	7	21.88
Prefer not to say	1	1.02	1	3.13	0	0	0	0
Gender								0
Male	44	44.9	13	40.63	14	41.18	17	53.13
Female	54	55.1	19	59.38	20	58.82	15	46.88
Ethnicity								
Asian	16	16.33	6	18.75	7	20.59	3	9.38
Black African/Caribbean	3	3.06	0	0	1	2.94	2	6.25
White	75	76.53	24	75	25	73.53	26	81.25
Mixed/multiple	1	1.02	1	3.13	0	0	0	0
Other/please specify	2	2.04	1	3.13	0	0	1	3.13
Prefer not to say	1	1.02	0	0	1	2.94	0	0
Region								
East Anglia	16	16.33	3	9.38	9	26.47	4	12.5
East Midlands	5	5.1	3	9.38	1	2.94	1	3.13
London	15	15.31	4	12.5	6	17.65	5	15.63
North West	6	6.12	1	3.13	0	0	5	15.63
Scotland	19	19.39	8	25	6	17.65	5	15.63
South East	7	7.14	3	9.38	2	5.88	2	6.25
South West	8	8.16	3	9.38	2	5.88	3	9.38
Wales	2	2.04	0	0	0	0	2	6.25
West Midlands	4	4.08	1	3.13	2	5.88	1	3.13
Yorkshire and the Humber	16	16.33	6	18.75	6	17.65	4	12.5

CPH: Clear Pass, Purple Hair. CPT: Clear Pass, Tattoo. CPA: Clear Pass, Accent.

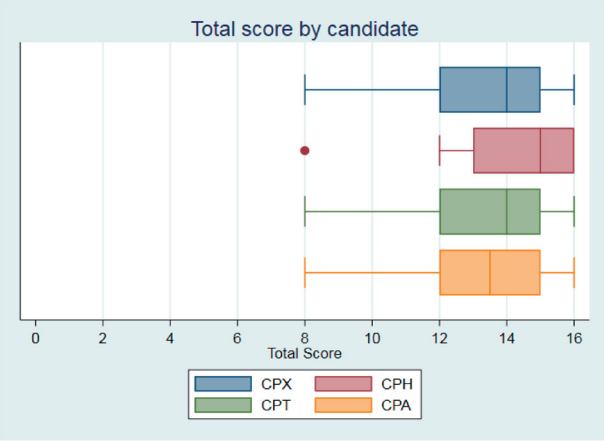

Figure 2. Total scores by candidate; CPX: Clear pass, no attribute, CPH: Clear pass, purple hair, CPT: Clear pass, tattoo, CPA: Clear pass, accent.

		CPX			
		Fail	Borderline	Clear Pass	Good
CPH	Fail	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Borderline	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Clear Pass	0 (0.0)	0 (0.0)	3 (9.4)	2 (6.3)
	Good	0 (0.0)	1 (3.1)	8 (25)	18 (56.3)

		CPX			
		Fail	Borderline	Clear Pass	Good
CPT	Fail	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Borderline	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Clear Pass	0 (0.0)	0 (0.0)	10 (29.4)	3 (8.8)
	Good	0 (0.0)	0 (0.0)	7 (20.6)	14 (41.2)

		CPX			
		Fail	Borderline	Clear Pass	Good
CPA	Fail	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Borderline	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Clear Pass	0 (0.0)	0 (0.0)	10 (31.3)	7 (21.9)
	Good	0 (0.0)	0 (0.0)	2 (6.3)	13 (40.6)

Figure 3. Number of assessors giving each combination of global grades to the candidate with no clear attribute (CPX) and the candidates with purple hair (CPH), tattoos (CPT), and an accent (CPA).

is in keeping with previous studies that suggest examiner bias is not responsible for the differential attainment amongst minority ethnic medical students (Yeates et al. 2017). The higher scores and global grades awarded to the candidate with purple hair may represent a positive contrast effect based on the presence of a notable characteristic which lead the candidate to stand out when compared to others (Yeates et al. 2015). However, any explanation for the difference in total scores is speculative at this stage and is likely to require further research.

The study used a randomised, single-blinded, controlled methodology to explore the influence of candidates' physical attributes on assessor ratings. However, the study does have some limitations. The study used video recordings of simulated performances and it is therefore possible that in real life assessors may be more or less vulnerable to bias than they were in this study. Further work should continue to explore the impact of bias in real-life assessments. We necessarily used different actors for each performance and

whilst every attempt was made to control for other sources of variability in the performances between candidates by standardising for age, gender and ethnicity and using a script, it is possible that minor variations between candidates persisted. All participants were volunteers and therefore it is possible that they are not a representative sample of the population of assessors as a whole. The study only explored the impact of physical attributes amongst white, female candidates performing at a clear pass standard and it is important to note that these findings may not be generalisable to candidates of other demographics, or to those performing at different levels. The study explored the impact of physical attributes in the context of an observed performance of a cranial nerve examination and we cannot exclude that different effects may occur in other types of assessment, particularly when they are more cognitively demanding for assessors. We also recognise the impact mark schemes and global grading systems may have on outcomes and our results may therefore not be generalisable if significantly different scoring rubrics are used. Further work is still needed to explore if other physical attributes such as choice of attire may have an impact on assessor ratings. It is also worth noting that any systematic effect of bias based on stereotype activation may vary over time as societal attitudes towards individual attributes also change.

Conclusion

Within the context of an online simulated assessment there does not appear to be any systematic effect of negative stereotype bias from assessors when rating competent candidates with tattoos, purple hair or a Liverpudlian accent when compared to a candidate without these characteristics.

Acknowledgements

The authors are grateful to all UK medical school assessment leads for their help in recruiting assessors. The authors are also grateful to the Medical Schools Council for administrative support with the study.

Disclosure statement

MG is supported by the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. CAB is supported by the NIHR Applied Research Collaboration (ARC) West Midlands. PY is funded by the NIHR Clinician Scientist Award. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Funding

The Medical Schools Council funded the recruitment of the simulated candidates, simulated patient and sourcing of the recording equipment for this study.

Notes on contributors

A. H. Sam, PhD, FRCP, SFHEA, is head of Imperial College School of Medicine and consultant physician and endocrinologist at Imperial College Healthcare NHS Trust.

M. D. Reid, MA, MB, BChir, MAcadMEd, MRCP, was a Clinical Education & Research Fellow at Imperial College London and is now a trainee in Geriatric Medicine at Kingston Hospital.

V. Thakerar, MBBS, MRCP, MRCGP, PGCert (MEd), is the lead for year 1 and 2 clinical placements at Imperial College School of Medicine and a practising general practitioner.

M. Gurnell, PhD, MA(MEd), FHEA, FAcadMEd, FRCP, is Clinical SubDean at the University of Cambridge School of Clinical Medicine and Professor of Clinical Endocrinology at Institute of Metabolic Science & Department of Medicine.

R. Westacott, MB, ChB, FRCP, is a Senior Lecturer in Medical Education at Birmingham Medical School and an acute medicine consultant at University Hospitals of Leicester NHS Trust (CCT in Nephrology).

P. Yeates, MRCP, PhD, is a senior lecturer in medical education research and a consultant in acute and respiratory medicine. His interests focus on assessor cognition and technology-enhanced assessment.

M. W. R. Reed, MD, BMedSci, MBCHB, FRCS, is a breast cancer surgeon who has been Dean of Brighton and Sussex Medical School since 2014 having moved from Sheffield University Medical School where he was head of Undergraduate Assessment for medicine. He is currently Co-Chair of Medical Schools Council and Chair of the education subcommittee.

C. A. Brown, PhD, SFHEA, is an Associate Professor in Quantitative Methods at Warwick Medical School. She has research interests in selection and assessment and teaches quantitative methods at all levels in Higher Education.

ORCID

A. H. Sam http://orcid.org/0000-0002-9599-9069
M. D. Reid http://orcid.org/0000-0002-8998-0265
M. Gurnell http://orcid.org/0000-0001-5745-6832
R. Westacott http://orcid.org/0000-0001-9846-1961
P. Yeates http://orcid.org/0000-0001-6316-4051
M. W. R. Reed http://orcid.org/0000-0001-7442-2132
C. A. Brown http://orcid.org/0000-0002-7526-0793

References

Beddow M. 2011. Hair color stereotypes and their associated perceptions in relationships and the workplace. [place unknown]; [accessed 2020 Feb 10]. https://pdfs.semanticscholar.org/57fd/6d85010f6db3475ee9acb9b651a9c6dcca27.pdf.

Bishop H, Coupland N, Garrett P. 2005. Conceptual accent evaluation: thirty years of accent prejudice in the UK. Acta Linguist Hafniensia. 37(1):131–154.

Chapman EN, Kaatz A, Carnes M. 2013. Physicians and implicit bias: How doctors may unwittingly perpetuate health care disparities. J Gen Intern Med. 28(11):1504–1510.

Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. 2009. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med. 24(1):74–79.

Deprez-Sims AS, Morris SB. 2010. Accents in the workplace: their effects during a job interview. Int J Psychol. 45(6):417–426.

Dovidio JF, Fiske ST. 2012. Under the radar: how unexamined biases in decision-making processes in clinical interactions can contribute to health care disparities. Am J Public Health. 102(5):945–952.

Gjerdingen DK, Simpson DE, Titus SL. 1987. Patients' and physicians' attitudes regarding the physician's professional appearance. Arch Intern Med. 147(7):1209–1212.

Gluszek A, Dovidio JF. 2010. The way they speak: a social psychological perspective on the stigma of nonnative accents in communication. Pers Soc Psychol Rev. 14(2):214–237.

Harasym PH, Woloschuk W, Cunning L. 2008. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Adv Health Sci Educ Theory Pract. 13(5):617–632.

Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. 2006. Assessing the mini-clinical evaluation exercise in comparison to a national specialty examination. Med Educ. 40(10):950–956.

Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. 2010. Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. Acad Med. 85(9):1453–1461.
Iramaneerat C, Yudkowsky R. 2007. Rater errors in a clinical skills assessment of medical students. Eval Health Prof. 30(3):266–283.
Johnston DW. 2010. Physical appearance and wages: do blondes have more fun? Econ Lett. 108(1):10–12.
Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. 2011. Opening the black box of clinical skills assessment via observation: a conceptual model. Med Educ. 45(10):1048–1060.
Macrae CN, Milne AB, Bodenhausen GV. 1994. Stereotypes as energy-saving devices: a peek inside the cognitive toolbox. J Pers Soc Psychol. 66(1):37–47.
McManus I, Thompson M, Mollon J. 2006. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ. 6(1):42.
Miller A, Archer J. 2010. Impact of workplace based assessment on doctors' education and performance: a systematic review. BMJ. 341(7775):c5064.
Moskowitz GB, Stone J, Childs A. 2012. Implicit stereotyping and medical decisions: unconscious stereotype activation in practitioners' thoughts about African Americans. Am J Public Health. 102(5):996–1001.
Norcini J, Burch V. 2007. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach. 29(9):855–871.
Ruetzler T, Taylor J, Reynolds D, Baker W, Killen C. 2012. What is professional attire today? A conjoint analysis of personal presentation attributes. Int J Hosp Manag. 31(3):937–943.
Stone J, Moskowitz GB. 2011. Non-conscious bias in medical decision making: what can be done to reduce it? Med Educ. 45(8):768–776.
Swanwick T. 2005. Informal learning in postgraduate medical education: from cognitivism to 'culturism'. Med Educ. 39(8):859–865.
Tavares W, Eva KW. 2014. Impact of rating demands on rater-based assessments of clinical competence. Educ Prim Care. 25(6):308–318.
Torre I, Goslin J, White L, Zanatto D. 2018. Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. In: ACM Int Conf Proceeding Ser. New York (NY): Association for Computing Machinery; p. 1–6
Williams RG, Klamen DA, McGaghie WC. 2003. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med. 15(4):270–292.
Wohlrab S, Stahl J, Kappeler PM. 2007. Modifying the body: motivations for getting tattooed and pierced. Body Image. 4(1):87–95.
Woolf K, Potts HWW, McManus IC. 2011. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. BMJ. 342(mar08 1):d901–d901.
Yeates P, Cardell J, Byrne G, Eva KW. 2015. Relatively speaking: contrast effects influence assessors' scores and narrative feedback. Med Educ. 49(9):909–919.
Yeates P, O'Neill P, Mann K, Eva K. 2013. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract. 18(3):325–341.
Yeates P, Woolf K, Benbow E, Davies B, Boohan M, Eva K. 2017. A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback and recollections in undergraduate clinical exams. BMC Med. 15(1):179.

Appendix 1

Video ordering

Version	Video 1	Video 2	Video 3	Video 4	Video 5
1	CPX	CPH	BL	CF	GD
2	CPX	CF	CPH	GD	BL
3	CPX	BL	GD	CPH	CF
4	CPX	GD	CF	BL	CPH
5	CPX	CPT	BL	CF	GD
6	CPX	CF	CPT	GD	BL
7	CPX	BL	GD	CPT	CF
8	CPX	GD	CF	BL	CPT
9	CPX	CPA	BL	CF	GD
10	CPX	CF	CPA	GD	BL
11	CPX	BL	GD	CPA	CF
12	CPX	GD	CF	BL	CPA

Key: CF: clear fail; BL: borderline; CPX: clear pass, no discernible attribute; CPH: clear pass, purple hair, CPT: clear pass, tattoo on both forearms; CPA: clear pass, regional accent; GD: good.